

Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures

Jifan Shi, Andrew E. Teschendorff, Weiyan Chen, Luonan Chen and Tiejun Li

Corresponding author: Andrew Teschendorff, CAS Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institute for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China, Department of Women's Cancer, UCL Elizabeth Garrett Anderson Institute for Women's Health, University College London, 74 Huntley Street, WC1E 6BT London, UK and UCL Cancer Institute, University College London, 72 Huntley Street, WC1E 6BT London, UK. E-mail: a.teschendorff@ucl.ac.uk; Luonan Chen, Key Laboratory of Systems Biology, CAS Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 200031 Shanghai, China, CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China, School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China. E-mail: lnchen@sibs.ac.cn; Tiejun Li, LMAM and School of Mathematical Sciences, Peking University, 100871 Beijing, China. E-mail: tieli@pku.edu.cn

Abstract

Motivation: Estimating differentiation potency of single cells is a task of great biological and clinical significance, as it may allow identification of normal and cancer stem cell phenotypes. However, very few single-cell potency models have been proposed, and their robustness and reliability across independent studies have not yet been fully assessed.

Results: Using nine independent single-cell RNA-Seq experiments, we here compare four different single-cell potency models to each other, in their ability to discriminate cells that ought to differ in terms of differentiation potency. Two of the potency models approximate potency via network entropy measures that integrate the single-cell RNA-Seq profile of a cell with a protein interaction network. The comparison between the four models reveals that integration of RNA-Seq data with a protein interaction network dramatically improves the robustness and reliability of single-cell potency estimates. We demonstrate that underlying this robustness is a correlation relationship, according to which high differentiation potency is positively associated with overexpression of network hubs. We further show that overexpressed network hubs are strongly enriched for ribosomal mitochondrial proteins, suggesting that their mRNA levels may provide a universal marker of a cell's potency. Thus, this study provides novel systems-biological insight into cellular potency and may provide a foundation for improved models of differentiation potency with far-reaching implications for the discovery of novel stem cell or progenitor cell phenotypes.

Key words: single-cell RNA-Seq; entropy; differentiation; potency; network; Waddington

Introduction

Over 60 years ago Waddington proposed an epigenetic landscape model of cellular differentiation, whereby cell fate transitions

are modeled as canalization events, with stable cell states defining the attractors or basins in this landscape [1, 2]. Focusing on a given cell lineage, a key ingredient of this landscape is the height of the attractor states [3], which is believed to correlate

Jifan Shi is a PhD student at the School of Mathematical Sciences, Peking University, Beijing, China.

Andrew Teschendorff is a Principal Investigator at the CAS-MPG Partner Institute for Computational Biology in Shanghai and a Royal Society Newton Advanced Fellow at University College London.

Luonan Chen is a Professor at the Key Lab for Systems Biology, Chinese Academy of Sciences, Shanghai, China.

Tiejun Li is a Professor at the School of Mathematical Sciences, Peking University, Beijing, China.

Submitted: 3 July 2018; Received (in revised form): 28 August 2018

© The Author(s) 2018. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

with differentiation potency, i.e. the relative number of downstream cell fate choices a given cell may have. Quantifying the differentiation potency of single cells is a task of critical importance, as it may help estimate potency heterogeneity within otherwise seemingly homogenous cell populations [4], it may facilitate explicit construction of Waddington landscapes [2, 5] and provide an unbiased means of identifying novel stem or progenitor cell phenotypes [6].

Given that mRNA expression levels of genes in a cell may inform on the activity of transcription factors and associated signaling pathways that control pluri-and-multi potency [7], it is reasonable to assume that differentiation potency is encoded by the genome-wide transcriptomic profile of the cell [6, 8–10]. Following this rationale, a number of explicit potency models based on biological principles have been proposed [6, 9–13]. For instance, StemID [6] estimates potency of a cell in terms of its genome-wide transcriptomic entropy, which measures how uniformly expressed the genes are. Because stemness is characterized by large numbers of differentiation expression programs being kept at a low and comparable level of basal activity, this transcriptomic entropy may therefore approximate potency. Another proposed model, called SLICE [9], groups genes into Gene-Ontology (GO) clusters and approximates potency of a cell in terms of a Shannon entropy, defined over gene expression derived activity estimates for each GO cluster: in effect, this measures the degree of uniformity of the GO-cluster activation levels and thus the degree of uniformity in the activation levels of different biological processes in a cell. A third model, called SCENT [10], integrates the single-cell RNA-Seq (scRNA-Seq) profile of a cell with a protein–protein interaction (PPI) network, modeling potency in terms of how efficient signaling can diffuse over the whole network. The efficiency of the diffusion process is estimated by the signaling entropy rate, which is expected to be lower for differentiated cells, since differentiation is accompanied by activation of a specific subnetwork, which in effect draws signaling flux away from other parts of the network [10]. Importantly, SCENT has been shown to estimate potency not only of normal cells, but also of cancer cells allowing putative cancer stem cell phenotypes to be identified [10].

Given the importance of estimating single-cell potency, a comparison of these different potency models is of great interest, and here we do so by performing a comprehensive comparison across a large number of independent scRNA-Seq datasets all profiling normal cells, in order to objectively determine which single-cell potency models are most accurate and robust. We do this in both single-cell as well as bulk-sample settings, while also adjusting for cell-cycle phase, which is a major confounder in scRNA-Seq studies. In performing these comparisons, we also introduce a novel single-cell potency measure, called Markov Chain Entropy (MCE), which like SCENT integrates the scRNA-Seq profile of a cell with a PPI network. Our results highlight that SLICE and StemID are less robust than SCENT/MCE in both single-cell and bulk tissue settings. We provide a detailed explanation and understanding as to the factors driving the increased robustness of MCE and SCENT, revealing that potency can be approximated in terms of the correlation between the transcriptome and connectome.

Materials & Methods

MCE

We here describe the construction of MCE to represent the potency of a sample with a scRNA-Seq or bulk RNA-Seq profile. The construction requires two inputs, i.e. the normalized RNA-

Seq profile and a fully connected signaling interaction network between the genes defined in the profile. For the network, we use the PPI network of 8434 proteins from [10]. We note that although in principle we could infer a gene regulatory network from the scRNA-Seq data itself using one of the many available methods [43–47], this does not work as well as using a PPI network, for reasons given later. Given a network/graph $G = (V, E)$ with n_V nodes in V and n_E edges in E , we model signaling as a particle on the network and let it do discrete-time Markov transition. Denote the transition probability matrix as $P = \{p_{ij}\}$, where $i, j = 1, \dots, n_E$. We set $p_{ij} = 0$ if there is no edge between two different nodes i and j . p_{ii} can be nonzero and is the probability of the particle staying at node i . We assume the network has a steady state with invariant distribution $\pi = (\pi_1, \dots, \pi_{n_V})$ after the particle walks for long enough time. As we shall see below, the invariant measure or steady-state probabilities will be determined from the sample's gene expression profile.

The *Markov Chain Entropy* of the Markov transition on network G is defined as

$$\text{MCE} = - \sum_{(i,j) \in \bar{E}} \pi_i p_{ij} \log(\pi_i p_{ij}), \quad (1)$$

where the summation is only for the terms $p_{ij} > 0$, and \bar{E} is the union set of E and all (i, i) self-loop edges. MCE can be split into two terms as

$$\text{MCE} = - \sum_i \pi_i \sum_{j \in \bar{N}(i)} p_{ij} \log p_{ij} - \sum_i \pi_i \log \pi_i, \quad (2)$$

where $\bar{N}(i)$ contains all the neighbors of node i including i itself.

The first term in Equation (2) is named as the *edge entropy* (similar to but not the same as the average local entropy/signaling entropy in [10, 42]), and the second term is called the *node entropy*. Since $\pi_i p_{ij}$ means the probability (or information flow or actual interaction) transmitted from i to j along the directed edge (i, j) in E , MCE is indeed the *whole entropy* of the information flow induced by the Markov chain on the graph or network G , rather than the *entropy rate* [10, 42]). Thus, it can be thought as a measurement of the heterogeneity of a network in terms of actual interactions. The more homogeneous of interactions the graph is, the bigger its MCE grows. We think that this intuition can be utilized to characterize the stemness of cells.

Computing MCE from single-cell RNA-seq dataset

Denote the dataset as $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}$, where n is the sample size, and $\mathbf{x}^{(i)}$ is a length- n_V column vector for gene expressions. n_V is not only the number of genes in consideration but also the counts of nodes in the gene network.

Computing MCE for each sample

Denote the 0–1 adjacency matrix of the PPI network as A , whose diagonal elements are also set to be 1s. For convenience, we still use X as the data matrix, and the size of A is $n_V \times n_V$. We have two assumptions before calculating the MCE for a single-cell sample \mathbf{x} :

- i. The normalized gene expression $\pi = \mathbf{x}/\|\mathbf{x}\|_{\mathcal{L}_1}$ is the invariant distribution of a discrete-time Markov transition matrix P on the gene network A ;
- ii. Among all the possible transition matrix P with invariant distribution π and topology A , we choose the one with the largest entropy.

That means the MCE problem can be formulated as

$$\begin{aligned} \text{MCE}(\mathbf{x}) = \max_{p_{ij} \geq 0} & - \sum_{(i,j) \in \bar{E}} \pi_i p_{ij} \log(\pi_i p_{ij}) \\ \text{subject to} & \sum_{j \in \mathcal{N}(i)} p_{ij} = 1, \quad i = 1, \dots, n_V, \\ & \sum_{i \in \mathcal{N}(j)} \pi_i p_{ij} = \pi_j, \quad j = 1, \dots, n_V, \end{aligned} \quad (3)$$

where the summation is only for the elements $(i, j) \in \bar{E}$ satisfying $A_{ij} = 1$, and $\mathcal{N}(i)$ includes node i and all its neighbors.

Because π is known (so is the node entropy), the above can be written into a standard convex optimization problem for edge entropy:

$$\begin{aligned} \min_{p_{ij} > 0} & \sum_{(i,j) \in \bar{E}} \pi_i p_{ij} \log(p_{ij}) \\ \text{s.t.} & P\mathbf{1} = \mathbf{1}, \\ & \pi P = \pi, \end{aligned} \quad (4)$$

where $P = (p_{ij})_{n_V \times n_V}$ has the same topology as A , and variables only are located at the position of 1s in A . Though the problem Equation (4) is convex with linear constraints, there are about $O(n_V^2)$ variables, which means to find an almost million-dimension optimizer for one cell! It sounds like infeasible for the first look, but luckily we can change the variables into its Lagrangian multipliers and solve it practically like

$$\begin{aligned} \mathcal{L}(p_{ij}, \lambda, \mu) = & \sum_{(i,j) \in \bar{E}} \pi_i p_{ij} \log(p_{ij}) - \sum_i \lambda_i \left(\sum_{j \in \mathcal{N}(i)} p_{ij} - 1 \right) \\ & - \sum_j \mu_j \left(\sum_{i \in \mathcal{N}(j)} \pi_i p_{ij} - \pi_j \right). \end{aligned} \quad (5)$$

Calculate the 1st derivations of (5) we get

$$\begin{aligned} \pi_i \log p_{ij} + \pi_i - \lambda_i - \pi_i \mu_j &= 0, \quad (i, j) \in \bar{E}, \quad i, j = 1, \dots, n_V, \\ \sum_{j \in \mathcal{N}(i)} p_{ij} &= 1, \quad i = 1, \dots, n_V, \\ \sum_{i \in \mathcal{N}(j)} \pi_i p_{ij} &= \pi_j, \quad j = 1, \dots, n_V, \end{aligned}$$

where only the terms $p_{ij} \geq 0$ corresponding to $A_{ij} = 1$ are used in the equations. Denote $\alpha_i = \exp(\lambda_i/\pi_i)$, $\beta_j = \exp(\mu_j - 1)$ and $p_{ij} = \alpha_i \beta_j a_{ij}$, then we have transformed the optimization problem into solving nonlinear equations.

$$\begin{cases} \alpha_i \sum_{j \in \mathcal{N}(i)} \beta_j = 1, & i = 1, \dots, n_V, \\ \beta_j \sum_{i \in \mathcal{N}(j)} \pi_i \alpha_i = \pi_j, & j = 1, \dots, n_V. \end{cases} \quad (6)$$

There are fortunately $2n_V$ nonlinear equations corresponding to $2n_V$ variables.

We propose an iteration method to solve the equations (6) as

1. $\alpha^{(n)} = 1/(A\beta^{(n-1)})$;
2. $\beta^{(n)} = \pi./(B'\alpha^{(n)})$,

where A is the adjacency matrix, B is got by multiplying each element in row i of A by π_i (i.e. $b_{ij} = \pi_i a_{ij}$), and $./$ means division by elements. With a proper initial point $\beta^{(0)}$, we can do the iteration until $(\alpha^{(n)}, \beta^{(n)})$ converging into a fixed point (α^*, β^*) .

Then the MCE of a single-cell sample can be calculated as

$$\begin{aligned} \text{MCE}(\mathbf{x}) = & - \sum_{(i,j) \in \bar{E}} \pi_i p_{ij} \log(\alpha_i^*) - \sum_{(i,j) \in \bar{E}} \pi_i p_{ij} \log(\beta_j^*) \\ & - \sum_i \pi_i \log \pi_i \\ = & - \sum_i \pi_i \log \alpha_i^* - \sum_j \pi_j \log \beta_j^* - \sum_i \pi_i \log \pi_i \\ = & - \sum_i \pi_i \log(\alpha_i^* \beta_i^* \pi_i). \end{aligned}$$

Normalization of MCE

To normalize the MCE into $[0, 1]$, we should compute the maximum possible MCE of all the networks with topology A . According to the formation of MCE in Equation (1), the maximum is reached when $\pi_i \sim d_i$ and $p_{ij} = 1/d_i$, where d_i is the degree of node i (including i itself). Denote $d = \sum_i d_i$ which is equal to the count of nonzero terms in A . Thus, we have

$$\max \text{MCE} = \log(d), \quad (7)$$

and the normalized MCE is computed as

$$\overline{\text{MCE}} = \frac{\text{MCE}}{\max \text{MCE}}. \quad (8)$$

Other entropy indices

Signaling entropy (SCENT)

Signaling entropy (SR) was proposed in [8, 10, 42]. It requires the specification of a gene functional network, e.g. a PPI network. The weights on the edges of this network are assumed to be proportional to the normalized expression levels of the genes, i.e. $w_{ij} \sim x_i x_j$. The transition matrix P has entries

$$p_{ij} = \frac{x_j}{\sum_{k \in \mathcal{N}(i)} x_k} = \frac{x_j}{(Ax)_i}, \quad (9)$$

where $\mathcal{N}(i)$ means the neighbors of node i and A is the PPI adjacency matrix. With detailed balance condition and invariant measure π computed from P as $\pi_i = x_i (Ax)_i / (\mathbf{x}' A \mathbf{x})$, signaling entropy is defined and calculated as

$$\text{Sr}(\mathbf{x}) = - \sum_i \pi_i \sum_{j \in \mathcal{N}(i)} p_{ij} \log p_{ij}. \quad (10)$$

To normalize the signaling entropy, maximum entropy is computed by

$$\max \text{Sr} = \log \lambda, \quad (11)$$

where λ is the largest eigenvalue of adjacency matrix A . Normalized entropy rate is defined as

$$\text{SR}(\mathbf{x}) = \frac{\text{Sr}(\mathbf{x})}{\max \text{Sr}}. \quad (12)$$

The main difference between SR and MCE are

- i. The adjacent matrix A has zero diagonal elements in SR, while it has ones in MCE. The basis of SR is interaction between proteins through the law of mass action, while the basis of MCE is a Markov transition on the gene network. The signaling particle in MCE has a probability to stay at one node which is more meaningful.

- ii. For MCE, we maximize the entropy to get the most likely transition matrix given a fixed invariant measure. In contrast, when computing SR, the weights of the stochastic matrix P are precomputed using the law of mass action, which is therefore fixed.

StemID

Grün et al. proposed the StemID algorithm to estimate differentiation potency of single cells in [6]. This algorithm does not use a PPI network. Strictly speaking it is also not a single-cell potency model, because it combines the computation of potency with a prior clustering step, which therefore uses information from all cells in the study. Specifically, StemID computes a stemness score s for each inferred cluster of cells as

$$\Delta E_k = \text{median}_{j \in k}(E_j) - \min_l(\text{median}_{j \in l}(E_j)), \quad (13)$$

$$s_k \propto \Delta E_k, \quad (14)$$

where k and l label clusters, and where E_j is the entropy of cell j defined as $E_j = -\sum_{i=1}^N p_{ij} \log_N p_{ij}$, with N the total number of genes and p_{ij} the normalized number of reads mapping to gene i in cell j . Thus, to compare fairly and objectively to other entropy indices like SR and MCE, we take E_j as the potency estimate from StemID for each cell. Importantly, we note that the StemID entropy is computed on the normalized (unlogged) counts, as recommended in the original paper [6].

SLICE

SLICE is another entropy-based index proposed in [9]. Like StemID it does not use a PPI network. Instead, it first clusters genes into m functional GO clusters, according to their similarity of GO annotation. For each cell j , SLICE then computes a probability activity score $p_{j,k}$ for GO-cluster k in cell j , by taking the average expression of genes mapping to a given GO-cluster k . Subsequently, it approximates potency as the Shannon entropy of the activity distribution, i.e. as $H_j = \mathbb{E}[-\sum_{k=1}^m p_{j,k} \log p_{j,k}]$, where the expectation is taken over a number of bootstraps. We computed SLICE as described in [10] using the R-functions provided by Guo et al. [9]. We note that, as with StemID, the computation of SLICE entropy is supposed to be done on the unlogged normalized counts. Thus, we used the downloaded normalized count data matrices when estimating SLICE entropy. The expression threshold parameter in SLICE was set to 1 (default value) for all studies, except for Yao1 where the data were already regularized (logged) and where we used a threshold of 0.1. We note that in the Yao1 study data have been renormalized by subsampling single cells to 20 000 unique molecular identifiers (UMIs) [14].

Single-cell and bulk RNA-Seq datasets

In total, we used scRNA-Seq datasets derived from nine independent studies. Two of these studies contained matched bulk RNA-Seq data. From all of these scRNA-Seq studies, we devised 11 independent analyses, comparing in each one cells of high potency to cells of low potency, assessing the different potency measures in their ability to discriminate these cell types. In the case of bulk RNA-Seq data, we devised five independent comparisons from the two studies with bulk RNA-Seq data. Table 1 gives a brief summary of the 11 independent comparisons performed on scRNA-Seq data and the five performed using bulk

RNA-Seq data. A more detailed description of the datasets is listed below:

- **Chu et al. datasets 1–4:** These datasets were derived from [38]. There are four experiments named Chu1, Chu2, Chu3 and Chu4. Chu1 dataset contains scRNA-Seq data for 1018 single cells, which is composed of 374 hESCs (human embryonic stem cells, 212 from H1 cell line and 162 from H9 cell line), 173 NPCs (neuronal progenitor cells, ectoderm derivatives), 138 DECs (definitive endoderm cells, endoderm derivatives), 105 ECs [endothelial cells, mesoderm progenitors (MPs)/derivatives], 69 TBs (trophoblast-like cells, extraembryonic derivatives) and 159 HFFs (human foreskin fibroblasts). The hESCs are pluripotent ($n = 374$) and the others are non-pluripotent ($n = 644$). Chu2 dataset is a time-course differentiation of single cells, in which hESCs were induced to differentiate into DECs, via a mesoendoderm intermediate. The time points cover before induction at $t = 0$ h ($n = 92$) and after induction at $t = 12$ h ($n = 102$), $t = 24$ h ($n = 66$), $t = 36$ h ($n = 172$), $t = 72$ h ($n = 138$) and $t = 96$ h ($n = 188$). There are 758 single cells in total. Chu3 dataset is from bulk samples. There are 19 bulk samples, which have seven samples for hESCs, two for NPCs, two for TBs, three for HFFs, three for ECs (MPs) and two for DECs. Chu4 is the time-course experiment of 15 bulk samples, consisting of three samples for each of five time points (12 h, 24 h, 36 h, 72 h and 96 h). Normalized data were downloaded from Gene Expression Omnibus (GEO) under accession number GSE75748 (files: GSE75748-bulk-cell-type-ec.csv, GSE75748-sc-cell-type-ec.csv, GSE75748-bulk-time-course-ec.csv, GSE75748-sc-time-course-ec.csv).
- **Trapnell et al. dataset:** This scRNA-Seq dataset was derived from [29]. It represents a time-course differentiation experiment of human myoblasts into human skeletal muscle cells. Human myoblasts were induced at $t = 0$ h ($n = 96$). Samples were also taken at $t = 24$ h ($n = 96$), $t = 48$ h ($n = 96$) and $t = 96$ h ($n = 84$) after induction. There are 372 single cells in total. Normalized data were downloaded from GEO under accession number GSE52529 (file: GSE52529-fpkmmatrix.txt).
- **Treutlein et al. dataset:** This dataset was derived from [39]. The experiment took samples from the developing mouse lung epithelium at embryonic days E14.5 ($n = 45$), E16.5 ($n = 27$), E18.5 ($n = 83$) and adulthood ($n = 46$), totalling 201 single cells. Normalized data were downloaded from GEO under accession number GSE52583 (file: GSE52583.Rda).
- **SemrauRB and SemrauSS datasets:** The two datasets named SemrauRB and SemrauSS were derived from [40], a study of retinoic acid (RA)-driven differentiation of pluripotent mouse embryonic stem cells (mESCs) to lineage commitment. SemrauRB used a recently developed Single Cell RNA Barcoding and Sequencing method (SCR-seq) and consists of 425 single-cell samples taken at nine time points during mouse embryonic differentiation. There were 53 single cells taken at 0 h. After all-trans RA exposure, 73 samples were taken at 6 h, 49 samples at 12 h, 27 samples at 24 h, 72 samples at 36 h, 43 samples at 48 h, 54 samples at 60 h, 29 samples at 72 h and 25 samples at 96 h. To increase power and because exit from pluripotency occurs after 12 h [40], all cells from 0–12 h were considered to be pluripotent. SemrauSS uses a different scRNA-Seq technology (SMART-seq). A total of 292 single cells were taken at four time points (75 samples at 0 h, 75 samples at 12 h, 81 samples at 24 h and 61 samples at 48 h). Normalized data were downloaded

Table 1. Summary of RNA-Seq experiments used. Table lists the name of the experiment, the type of RNA-Seq data (if single cell or bulk), the species, the total number of profiles (after quality control), the time points or potency states being compared, the cell types being compared, the number of single cells in each comparison group and the reference. Chu1–4 derive from Chu *et al.*, Trapnell derives from Trapnell *et al.*, Treutlein from Treutlein *et al.*, SemrauRB & SemrauSS derive from the RNA Barcoding (RB) and Smart-Seq (SS) scRNA-Seq experiments in Semrau *et al.*, WangNPC derives from Wang *et al.*, YangML derives from Yang *et al.* and Yao1 derives from a scRNA-Seq experiment that used UMIs, while Yao2 is a corresponding matched bulk RNA-Seq dataset. Abbreviations: Pl, pluripotent; NonPl, non-pluripotent; E, embryonic day; h, hours; d, day, MP, multipotent progenitor; NPC, neural progenitor cell. For instance, for Chu1 set, we tested the alternative hypothesis that the potency measure is higher in the pluripotent cells compared to non-pluripotent cells, and of the 1018 high-quality scRNA-Seq profiles, 374 were pluripotent and 644 were non-pluripotent

Name	Type	Species	Total number (after QC)	Comparison	Cell types in comparison	Numbers in comparison	Ref
Chu1	scRNA	Human	1018	Pl > NonPl	hESCs to MPs	374 versus 644	[38]
Chu2	scRNA	Human	758	0 h > 96 h	hESCs to MPs	92 versus 188	[38]
Trapnell	scRNA	Human	372	0 h > 72 h	Myoblasts to skeletal muscle	96 versus 84	[29]
Treutlein	scRNA	Mouse	201	E14 > Adult	Lung Differentiation	45 versus 46	[39]
SemrauRB	scRNA	Mouse	425	0–12 h > 96 h	hESCs to MPs	175 versus 25	[40]
SemrauSS	scRNA	Mouse	292	0 h > 48 h	hESCs to MPs	75 versus 61	[40]
WangNPC	scRNA	Human	483	0–1 d > 30 d	NPCs to Neurons	158 versus 81	[17]
YangML	scRNA	Mouse	447	E10 > E17	Liver Differentiation	54 versus 70	[16]
Yao1a	scRNA	Human	2684	0 d > 26 d	hESCs to NPCs	40 versus 595	[14]
Yao1b	scRNA	Human	2684	0 d > 54 d	hESCs to Neurons	40 versus 765	[14]
Yao1c	scRNA	Human	2684	26 d > 54 d	NPCs to Neurons	595 versus 765	[14]
Chu3	bulkRNA	Human	19	Pl > NonPl	hESCs to MPs	7 versus 12	[38]
Chu4	bulkRNA	Human	15	12/24 h > 72/96 h	hESCs to definite endoderm	6 versus 6	[38]
Yao2a	bulkRNA	Human	33	0–9 d > 26 d	hESCs to NPCs	12 versus 6	[14]
Yao2b	bulkRNA	Human	33	0–9 d > 54 d	hESCs to Neurons	12 versus 4	[14]
Yao2c	bulkRNA	Human	33	26 d > 54 d	NPCs to Neurons	6 versus 4	[14]

from GEO under accession number GSE79578 (files: GSM2098545-scrbseq-2i.txt, GSM2098550-scrbseq-48h.txt, GSM2098555-smartseq-12h.txt, GSM2098546-scrbseq-6h.txt, GSM2098551-scrbseq-60h.txt, GSM2098556-smartseq-24h.txt, GSM2098547-scrbseq-12h.txt, GSM2098552-scrbseq-72h.txt, GSM2098557-smartseq-48h.txt, GSM2098548-scrbseq-24h.txt, GSM2098553-scrbseq-96h.txt, GSM2098549-scrbseq-36h.txt, GSM2098554-smartseq-2i.txt).

- Wang *et al.* dataset: This dataset was derived from [17], a study of non-directed differentiation over a 30-day period of neural progenitor cells (NPCs) into developing neurons, with the NPCs derived from hESCs. After quality control, there were 483 usable single cell samples describing the differentiation from NPCs (at day 0) to neurons and other differentiated cell-types. Cell numbers were 80 at day 0, 78 at day 1, 85 at day 5, 80 at day 7, 79 at day 10 and 81 at day 30. Normalized data were downloaded from GEO under accession number GSE102066 (file: GSE102066-normalized-counts.experiment1.DataMatrix.txt).
- Yang *et al.* dataset: This scRNA-Seq dataset was derived from [16], a study of differentiation of mouse hepatoblasts into hepatocytes and cholangiocytes. After quality control, 447 single-cell samples were taken during embryonic development. There are 54 single cells at embryonic day 10.5 (E10.5), 70 at E11.5, 41 at E12.5, 65 at E13.5, 70 at 14.5, 77 at 15.5 and 70 at E17.5. Normalized data were downloaded from GEO under accession number GSE90047 (file: GSE90047-Single-cell-RNA-seq-TPM.txt).
- Yao *et al.* datasets 1–2: This scRNA-Seq dataset (Yao1) was obtained by using a method based on multiplexed single-cell RNA-seq (CelSeq) and UMIs [14]. Cel-Seq was used to profile cells at multiple time points (days 0, 12, 19, 26, 40 and 54) in a study of differentiation of hESCs into cortical and non-cortical progenitor and finally into neurons.

After quality control, there were 2684 single cells: 40 hESCs, 504 collected at day 12, 278 at day 19, 595 at day 26, 502 at day 40 and 765 at day 54. Progenitor expansion into neurons occurred at day 26. Normalized data were downloaded from GEO under accession number GSE86977 (file: GSE86977-UMI-20K.2684.csv). In addition, we also analyzed a matched bulk RNA-Seq dataset (Yao2) from the same study, encompassing a total of 33 samples: five hESC samples (day 0), three samples at day 6, four samples at day 9, six samples at day 12, five samples at day 19, six samples at day 26 and four samples at day 54. Normalized data were downloaded from GEO under accession number GSE86985 (file: GSE86985-truseq.tpm.csv).

Further processing and quality control analysis

We used the following general procedure to further assess quality and further normalize the provided scRNA-Seq data. First, in all cases, we computed the coverage per single cell, i.e. the number of genes with normalized counts (e.g. transcripts per million (TPM) fragment reads per kilobase exon per million (FPKM)) above zero. Any single cell with less than 10% coverage was removed. Next, we quantile normalized (QN) the resulting dataset to remove potential batch effects. For most datasets, we then renormalized the minimum read count to 1, i.e. any normalized count less than 1 was renormalized to 1. We justify this on the basis that the normalized counts of 1 and 0 are effectively indistinguishable. The threshold of 1 was derived by inspection of the distribution of normalized counts. After QN, we log₂-transformed the data with an offset value of 0.1 added before log-transformation. The offset was added in order to ensure that the minimum value after log-transformation would not be 0, but a nonzero value (typically log₂ 1.1 ≈ 0.13). We did not want zeroes in our final matrix to avoid singularities in

the computation of the signaling entropy rate used in SCENT. The log-transformation was done to stabilize the variance and regularize the effects of outliers. Next, in the case of human data, we mapped gene symbols to Entrez gene IDs (if not annotated to Entrez gene IDs already) and finally averaged any entry mapping to the same Entrez gene ID. In the case of mouse, we first obtained the human homologs using the *annotationTools* Bioconductor package (<http://www.bioconductor.org>) and finally mapped to human Entrez gene IDs. The typical range of a resulting data matrix was then typically 0.13 to approximately 16.

We note that the Yao1 scRNA-Seq dataset used CelSeq-technology and UMIs and was normalized using a different procedure as described in [14]. For this dataset, we therefore used a slightly different choice of thresholds. First, the data downloaded was already on a log-scale, but still with a few significant outliers: thus, we regularized the data further by setting any values larger than 16 to 16. We then performed QN and finally any value less than 1 was set to a value of 0.1 (in order to remove zeroes from the data matrix).

The above procedure resulted in data matrices defined over human Entrez gene IDs and single cells of following dimensions: Chu1 (18 935 genes, 1018 single cells), Chu2 (18 935 genes, 758 single cells), Chu3 (18 935 genes, 19 bulk samples), Chu4 (18 935 genes, 15 bulk samples), Trapnell (22 351 genes, 372 single cells), Treutlein (12 413 genes, 201 single cells), SemrauRB (13 755 genes, 425 single cells), SemrauSS (13 663 genes, 292 single cells), WangNPC (8242 genes, 483 single cells), YangML (16 119 genes, 447 single cells), Yao1 (22 113 genes, 2684 single cells) and Yao2 (19 659 genes, 33 bulk samples).

Evaluation framework

In order to compare the different potency measures (described in the next subsections), we assessed their ability to discriminate single cells (and bulk samples) that ought to differ substantially in terms of their potencies. We reasoned that potency measures which would fail to discriminate such cell types would be less preferable than those that are more robust. For each of the studies, we therefore devised comparisons between cell types (or bulk samples) where typically we would expect to see a big difference in potency, so that it is reasonable to assume that higher Area Under the Curves (AUCs) (discriminatory ability) and more significant *P*-values (as derived using one-tailed Wilcoxon rank sum tests) would indicate better potency measures. For the scRNA-Seq datasets, we devised a total of 11 independent comparisons, where the specific cell types being compared and the corresponding cell numbers are given in Table 1. Below, we briefly summarize the comparisons.

In Chu1, we chose to compare the pluripotent hESCs to all progenitor and non-pluripotent cell types. In Chu2, we compared the hESCs at time 0 to the final time point (96 h) corresponding to definite mesoderm/endoderm (i.e. non-pluripotent). For the corresponding bulk sets Chu3 and Chu4, we chose the same groups, although for Chu4 only 12 h and 24 h were available to represent the pluripotent state, so for Chu4 we compared 12 h/24 h to 72 h/96 h, which yielded six samples in each group. Only using 12 h and 96 h would result in too few samples. We also note that the transition to the definite mesoendoderm occurs at 72 h and beyond. In Trapnell, we compared the myoblasts at 0 h to the differentiated skeletal muscle cells at 72 h. In Treutlein, we compared embryonic day E14 to the adult differentiated cells. In SemrauRB, because exit from pluripotency occurs after 12 h and because cell numbers at each time point were fairly low, we compared all cells harvested at 0, 6 and 12 h to those harvested

at 96 h. In SemrauSS, we compared cells at 0–48 h. In WangNPC, we compared neural progenitors at 0 and 1 day to differentiated neural cells at 30 days. We deemed safe to bunch together 0 and 1 day cells as differentiation occurred much later in the assay and to increase cell numbers for the more potent NPC phenotype. In YangML, we compared embryonic day 10 (E10) to E17 single cells. In Yao1, we devised three separate comparisons, because in this study there were many single cells and because the nature of the experiment allowed precise definition of three potency states: hESCs at day 0, cortical and non-cortical progenitors at day 26 and differentiated neurons at 54 days. Thus, in Yao1a we compared day 0 to day 26, in Yao1b, we compared day 0 to day 54, and in Yao1c we compared day 26 to day 54. For the corresponding bulk sample study, Yao2, we also performed three comparisons, but in order to increase sample numbers we considered the pluripotent state to also include day 9 samples, which is justified based on the findings of the original publication [14].

Adjusting for cell-cycle phase

Differences in cell potency between groups could be confounded by differences in the proportion of cycling versus non-cycling cells. For each single cell we thus computed two cell-cycle scores (one for the G1-S and the other for the G2-M phase), using the procedure described in [10, 41]. Because the relation between potency and cell-cycle scores could be highly nonlinear (as described in [10]), linear adjustment procedures are inappropriate. Another approach to compare potency of cell types, adjusted for cell cycle, would be to set a threshold on the scores and only use cells with cell-cycle scores below a given threshold. However, these thresholds may be study-specific, and we found that if we were to use the same threshold for all studies, that this may lead for specific studies to highly unequal numbers of low-cycling cells in each comparison group or to effectively no cells in a given group. For instance, in some studies we find that all single hESCs exhibit relatively high-cycling scores. This indicates that disentangling cell-cycle effects and potency is extremely difficult, if not impossible, and may also be strongly dependent on the experimental details (e.g. culture conditions of cells, etc.). Thus, we decided on a different approach which would allow us to compare cell types, while adjusting for cell-cycle phase as much as possible. Specifically, for each study and cell type of interest, we chose a lower quantile of cell-cycle scores based on the scores of all cells within that cell type. Doing so, ensure that we are comparing sufficient numbers of single cells within each group, but also that we are using the lowest-cycling cells within each group. The quantiles were chosen to ensure sufficient numbers of cells in each comparison group. We note that since there are two cell-cycle scores, which we performed this operation on G1-S and G2-M separately, subsequently only using the single cells that were in the lower-defined quantile for both phases. Thus, absolute cell numbers could be low for some studies, but with the chosen quantiles enough cells were being compared to achieve potential statistical significance. The quantiles chosen, the number of low-cycling single cells in each comparison group and the results obtained are listed in Table S2.

Construction of a model landscape

We first use t-stochastic neighborhood embedding (t-SNE) [48] to dimensionally reduce and visualize the single-cell data, projecting the data onto a lower two-dimensional space. The t-SNE coordinate values are used to model the *x* and *y* axes. MCE is used to represent the value on the *z*-axis. In the tSNE-MCE

figure, cells located higher up along the z-axis are more potent as they have higher entropy. To plot an illustrative potential landscape, we use a function like

$$f(x, y) = 1 - \frac{4-y}{4} \left(e^{-80(x-\frac{1}{8})^2} + e^{-40(x-\frac{3}{8})^2} + e^{-60(x-\frac{5}{8})^2} \right) - \frac{y}{4} e^{-40(x-\frac{9}{16})^2},$$

$$h(x, y) = 8f(x, y) + y,$$

and plot the height values of $h(x, y)$ for $(x, y) \in [0, 1] \times [-1, 4]$. The idea of constructing such landscape function is to make it a Gaussian mixture in x-direction but linear in y-direction. Finally, we place the surface at the right height in the tSNE-MCE figure. We point out that the resulting surface map is only intended as an illustration of Waddington's landscape to merely show how stem cells differentiate into multipotent cells from the three main germ layers.

Results

Modeling single-cell potency as the graph entropy of a Markov chain process

The measured mRNA expression profile of a cell can be interpreted as the net result of a complex signaling process that takes place in the cell. Assuming an interaction network (e.g. a PPI network), one may thus attempt to model a given signaling process in terms of a Markov chain that is consistent with the observed expression profile of the cell (Methods, Figure 1A).

For each Markov-chain process one can further define an associated graph entropy, here termed MCE, that measures the overall uniformity of the signaling fluxes in the network (Methods, Figure 1A). We posited that among all Markov chains consistent with the observed expression profile, the one maximizing the graph entropy could provide us with an estimate of single-cell potency (Methods, Figure 1A). To understand this, we note that higher entropy means a more uniform Markov chain signaling process on the network, allowing signaling information to diffuse more efficiently. Thus, since a terminally differentiated cell (a cell of lowest potency) is characterized by the activation of specific signaling subnetworks (i.e. signaling pathways), it is plausible that this nonuniformity can be detected in terms of a lower MCE.

Comparison of single-cell potency measure reveals robustness of MCE and SCENT

To compare the different single-cell potency models we devised an evaluation framework, designed to be as objective as possible: we reasoned that although a population of seemingly equally potent single cells may not be entirely so due to inter-cellular heterogeneity a reliable potency measure should be able to nevertheless discriminate single-cell populations that differ significantly in terms of differentiation potency. For instance, a robust single-cell potency measure should be able to discriminate pluripotent from non-pluripotent cell populations or cells at the start and endpoints in time-course differentiation experiments. Thus, we assembled a collection of nine scRNA-Seq experiments, encompassing a total of approximately 6700 high-quality scRNA-Seq profiles, from which we devised 11 independent comparisons between cell types that ought to differ significantly in terms of differentiation potency (Methods, Table 1).

To each of these datasets, we applied the three previously proposed single-cell potency models (StemID [6], SLICE [9] and SCENT [10]), as well as MCE.

We first considered a time-course differentiation experiment in which hESCs were induced to differentiate into cortical NPCs and finally into neurons over a 54 day period, with scRNA-Seq measurements taken at six time points, encompassing a total of 2684 single cells [14]. While MCE and SCENT showed the expected gradual decrease of potency with differentiation time point, StemID exhibited a less significant pattern, while SLICE did not exhibit a decrease (Figure 2A).

Comparing hESCs to NPCs at day 26 (Figure 2A, Yao1a in Table 1) and separately hESCs to differentiated neurons at day 54 (Figure 2A, Yao1b in Table 1) revealed that SCENT and MCE could discriminate cells with almost perfect accuracy (see Yao1a and Yao1b in Figure 2B, Supplementary Table S1). As expected the difference in potency between NPCs and neurons (Yao1c in Table 1) was less strong but still significantly stronger than the differences predicted by StemID or SLICE (see Yao1c in Figure 2B, Supplementary Table S1). As assessed over the 11 independent comparisons (Table 1, Supplementary Figure S1), we observed that MCE and SCENT outperformed the other two (SLICE and StemID), also when adjusted for cell-cycle phase (Figure 2B and C, Supplementary Tables S1 and S2, Methods). In 9/11 comparisons, MCE and SCENT were better than StemID, and in all 11 better than SLICE. The better performance of MCE and SCENT was statistically significant under a paired Wilcoxon test over the 11 comparisons (Figure 2B and C).

MCE and SCENT outperform other potency models on bulk RNA-Seq data

Since potency is also a property of a cell population [10, 15], any robust measure of single-cell potency should, in principle, also work for bulk samples since a bulk RNA-Seq profile is effectively an average over single cells [8]. We observed that only SCENT and MCE exhibited a robust performance on bulk samples (Supplementary Table S3), in line with the results obtained on scRNA-Seq data. Thus, based on the comparative analysis performed here, we conclude that SCENT and MCE currently provide the more robust measures of cell potency, and therefore advocate their use for explicit, unbiased construction of Waddington landscapes. For instance, we applied MCE to the Chu1 dataset (Table 1, Methods), in combination with t-SNE [48], to construct a Waddington epigenetic landscape encompassing four attractor states, representing pluripotent hESCs and multipotent (i.e. non-pluripotent) progenitors from the three main germ layers: ectoderm, mesoderm and endoderm (Figure 1B).

MCE correctly predicts differentiation time points

To further validate MCE we applied it to two different time-course differentiation experiments, generating three-dimensional representations of cell-lineage trajectories. For an embryonic time-course differentiation experiment of mouse hepatoblasts into hepatocytes and cholangiocytes (Yang et al., Table 1), MCE predicted separation after embryonic day 13 (Figure 3A), consistent with the findings of [16]. When applied to a 30 day time-course differentiation experiment of neural progenitor cells into neurons (Wang et al., Table 1), MCE correctly predicted a gradual decrease in potency (Figure 3B, Supplementary Figure S1).

Thus, these applications illustrate that the estimate of cell potency as given by MCE is able to track gradual changes in

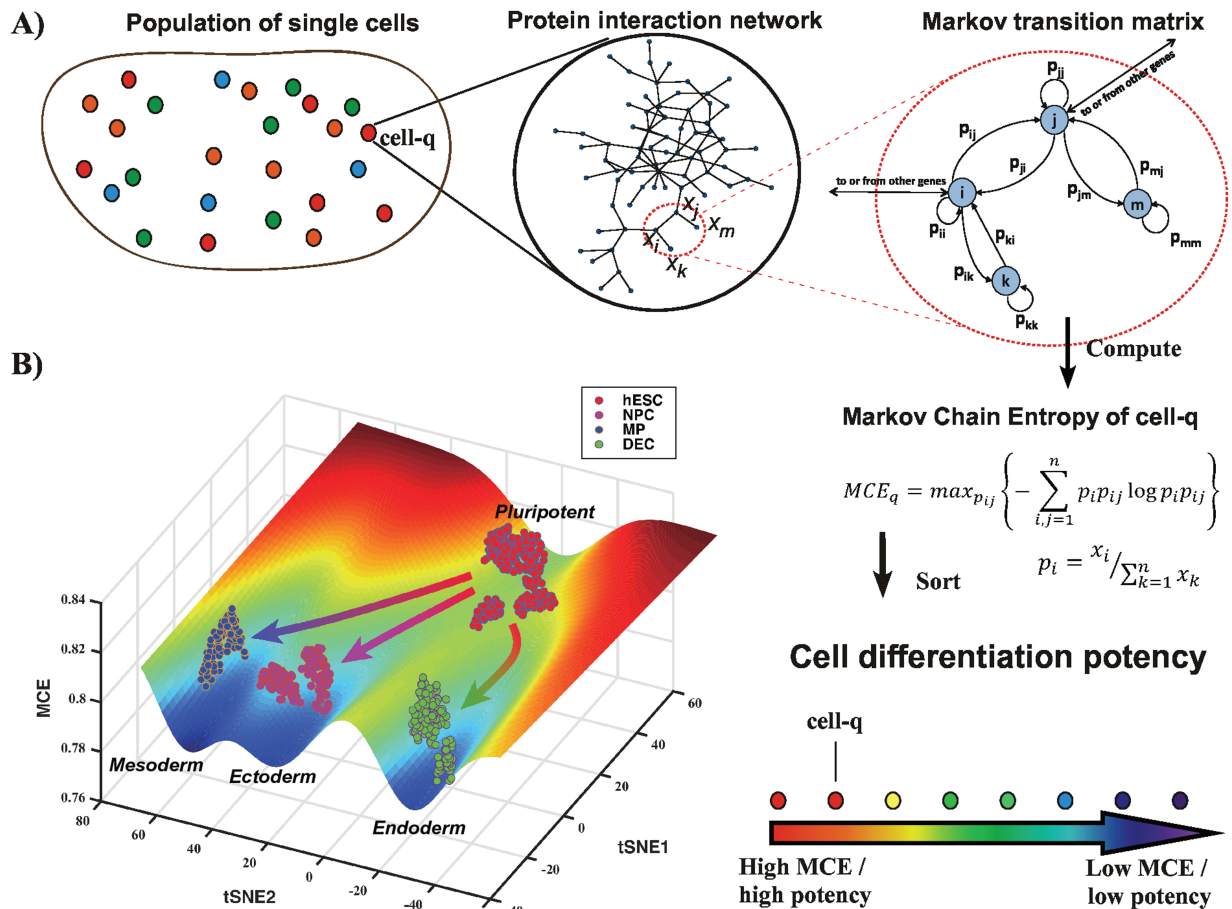


Figure 1. Markov-Chain Entropy (MCE) for quantifying single-cell potency in Waddington's landscape. (A) The key assumption underlying the estimation of MCE is that the measured gene expression profile $\mathbf{x} = (x_1, \dots, x_n)$ of a cell q is the end result of an unknown Markovian signaling process in the cell, which however we can try to estimate by using a network model of signaling interactions. This network imposes constraints on the space of all possible Markov transition matrices, and MCE is estimated from the one maximizing the Shannon entropy of the Markov process. Having estimated MCE for each single cell, we can then sort them according to their MCE, which we posit is a proxy for cell potency. (B) An example of a Waddington epigenetic landscape for the Chu1 et al. dataset (Table 1), depicting four main attractor states, representing hESCs (pluripotent), NPCs (ectoderm, multipotent), MPs (mesoderm, multipotent) and definite endoderm progenitor cells (DECs, endoderm, multipotent). The z-axis labels the MCE values of over 1000 single cells representing these four cell types, whereas the y-axis and x-axis label the projections inferred using t-SNE.

potency with time, as well as specific time points at which differentiation into distinct cell fates occur.

Cellular potency encoded by the Pearson correlation of transcriptome and connectome

In order to understand the increased robustness of MCE and SCENT over StemID and SLICE, we note that both MCE and SCENT integrate the scRNA-Seq data with orthogonal (i.e. completely independent) connectivity information as derived from a PPI network [10] (Figure 1A, Methods). While such networks are mere caricatures of the complex spatially and temporally dependent signaling processes that take place in a cell, it is plausible that there are specific robust network features that could contribute to the observed robustness of the MCE and SCENT measures. For instance, the difference in connectivity between a hub and a low-degree node is likely to be true irrespective of the PPI network considered and also likely to hold in a wide range of different cell types and biological contexts where both proteins are expressed at similar levels. To explore this, we first verified that the network is indeed a critical feature underlying

the robustness of MCE as a potency measure. Specifically, using the neural progenitor to neuron differentiation experiment from Wang et al. [17], we observed how the discrimination accuracy of MCE decreased significantly upon permuting the expression values over the nodes in the network (Figure 4A–C).

Since the connectivity (or degree) profile is one of the most fundamental properties of a network, we posited that MCE might be approximated by the Pearson correlation between the transcriptome and connectome. We verified that approximately 70% of the variance in MCE can be explained by a linear correlation between transcriptome and connectome (Figure 4D). In line with this, we observed that t-statistics of differential expression between NPCs and differentiated neurons were generally positive (i.e. higher expression in NPCs compared to neurons) for network hubs, while low-degree nodes did not exhibit any preference for over or underexpression (Figure 4E). We verified that all of these patterns were a systemic feature, observed in effectively all major datasets analyzed here (Supplementary Figures S2–S10). Specifically, (i) the discrimination accuracy of MCE drops significantly upon permuting expression values over the network nodes, confirming the importance of the network;

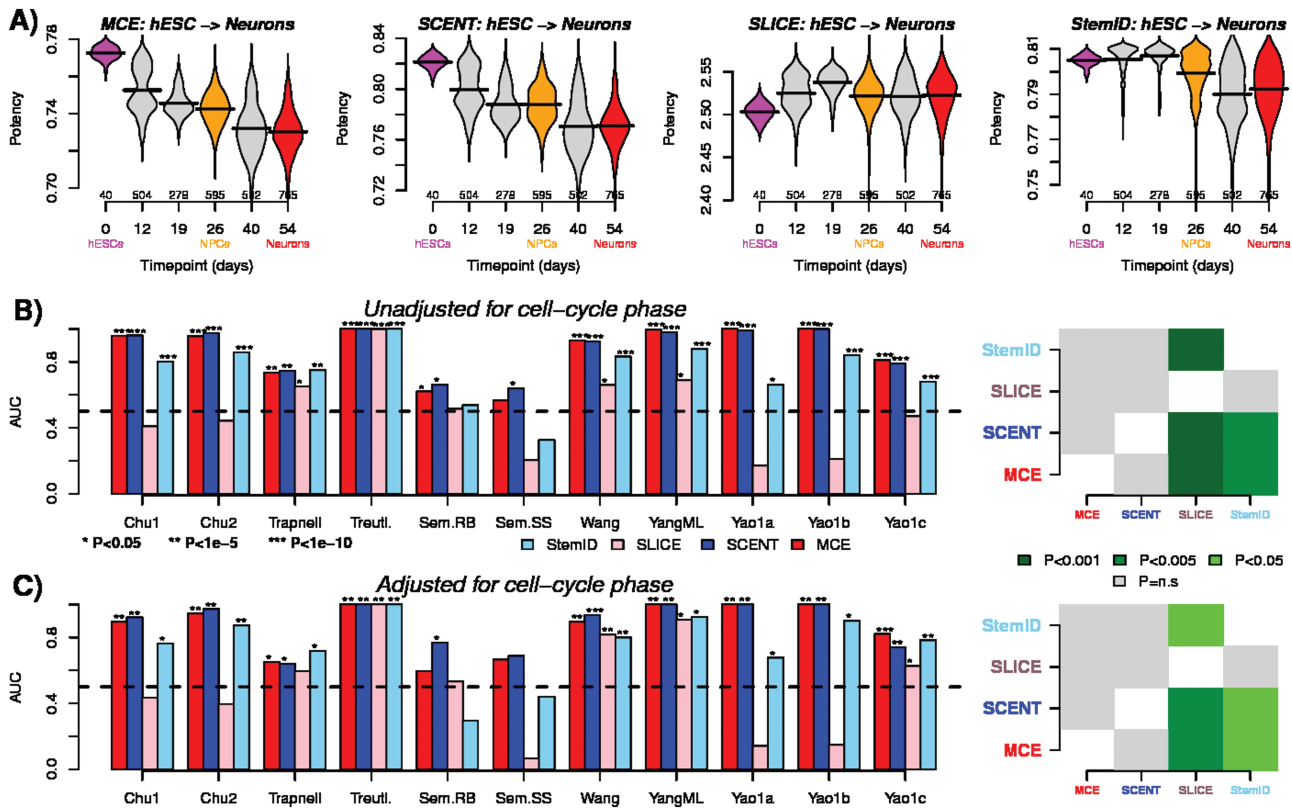


Figure 2. Comparison of discrimination accuracy of single-cell potency measures across scRNA-Seq datasets. (A) For the scRNA-Seq dataset (Yao et al. [14]), which profiled over 2600 cells in a time-course differentiation of hESCs (day 0) to NPCs (day 26) and finally to terminally differentiated neurons (day 54), we show corresponding violin plots of the predicted potencies at each measured time point and for the four different potency measures (SLICE, StemID, MCE and SCENT), as indicated. The number of cells at each time point is given below violins. (B) **Left panel:** Barplots compare the AUCs of the four different single-cell potency measures across 11 independent comparisons drawn from nine independent scRNA-Seq experiments (see Table 1 for definitions of cell types being compared). The AUCs reflect the measure's discrimination accuracy of single cells that ought to differ in terms of differentiation potential and can be derived from the statistic of a Wilcoxon rank sum test. P-values are also derived from a one-tailed Wilcoxon rank sum test comparing cell types. **Right panel:** Heatmap of meta-analysis P-values comparing single-cell potency measures to each other. P-values estimated from a paired one-tailed Wilcoxon rank sum test over the 11 independent comparisons. In the heatmaps, the P-value entry for column *i* and row *j* is for testing the alternative hypothesis that method in row *j* has higher AUC values than method in column *i*. (C) As (B), but now adjusting for cell-cycle phase, i.e. using only the cells with the lowest-cycling scores in each comparison group.

(ii) the Pearson correlation between transcriptome and connectome explains a significant proportion of the variance in MCE; and (iii) differentiation potency is encoded by a positive correlation between transcriptome and connectome with network hubs always preferentially overexpressed in the cells of higher potency (Supplementary Figures S2–S10).

Biological significance of overexpressed network hubs

In order to gain a better understanding of the above phenomenon, we performed a Gene-Set-Enrichment-Analysis (GSEA) on the network hubs overexpressed in more potent cells. Specifically, we considered the comparison between hESCs and NPCs, as there were two independent scRNA-Seq datasets available (Chu1, Yao1 and Table 1) to assess overall robustness of our findings. Focusing first on Yao1 et al., we verified that network hubs do indeed exhibit preferential overexpression in hESCs compared to NPCs (Figure 5A), a result also evident in Chu1 et al. (Figure 5B).

Performing differential expression analysis using an empirical Bayes framework [18], we next ranked all network hubs (defined as those with a degree ≥ 90) according to overexpression in hESCs and then again by underexpression. Of the

1406 hubs, 376 and 102 were significantly overexpressed and underexpressed (adjusted P-value < 0.05), respectively. Performing GSEA [19] on the overexpressed subset revealed a striking and highly significant enrichment for genes encoding stemness, glutamine metabolism and mitochondrial ribosomal proteins (Supplementary Table S4) with top-ranked biological terms highly reproducible between the two independent scRNA-Seq datasets (Figure 5C). This enrichment was only observed for overexpressed network hubs and not for underexpressed ones (Supplementary Figure S11, Supplementary Table S5). Corresponding heatmaps of gene expression confirmed the overexpression of mitochondrial ribosomal genes in both scRNA-Seq datasets (Figure 5D). These findings are highly consistent with those of a recent study, which showed that suppression of transcription of ribosomal genes and upregulation of lineage-specific factors in hematopoiesis coordinately control lineage differentiation [20].

Discussion

We have here performed a comprehensive comparison between four different single-cell potency models. All four potency models use a notion of entropy to quantify differentiation potency

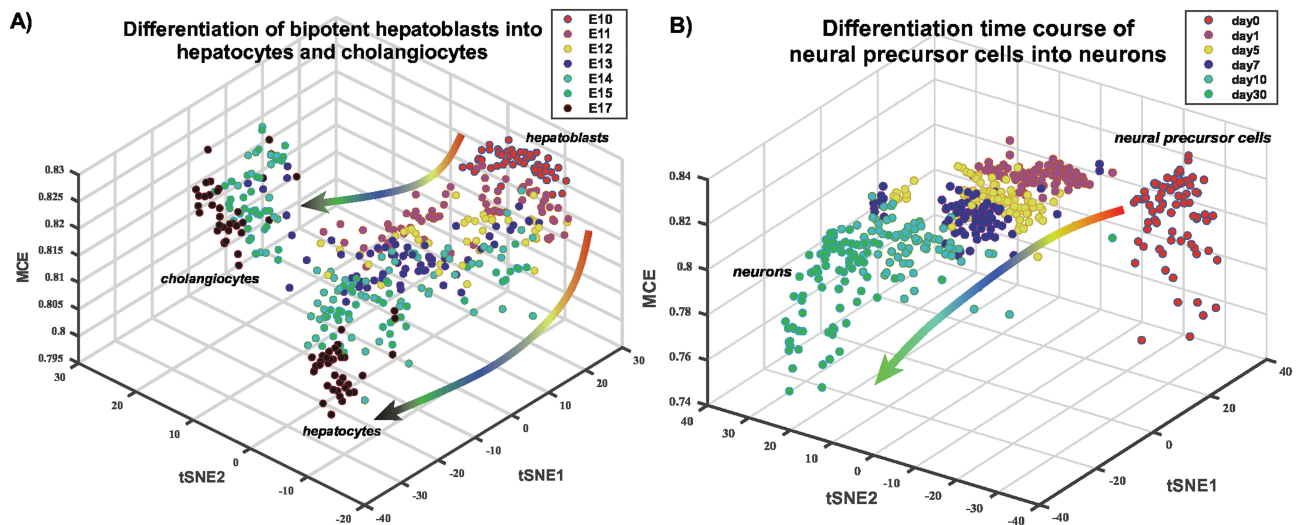


Figure 3. Three-dimensional representation of single-cell differentiation trajectories according to MCE. (A) Differentiation landscape of hepatoblasts. Figure depicts a three-dimensional representation of single cells in an embryonic differentiation time course of mouse hepatoblasts into hepatocytes and cholangiocytes (Yang et al.), as indicated. The MCE potency estimate labels the z-axis, whereas the x and y axes correspond to the tSNE components. Observe how the model predicts separation/differentiation into hepatocytes and cholangiocytes at embryonic day 13 (E13) in line with the results of Yang et al. (B) Differentiation landscape of neural precursor cells into neurons. Figure depicts a three-dimensional representation of single cells in a differentiation time course of human neural precursor cells (day 0) into neurons (day 30) (Wang et al.), as indicated. Observe how MCE correctly predicts a gradual decrease in potency with time of differentiation.

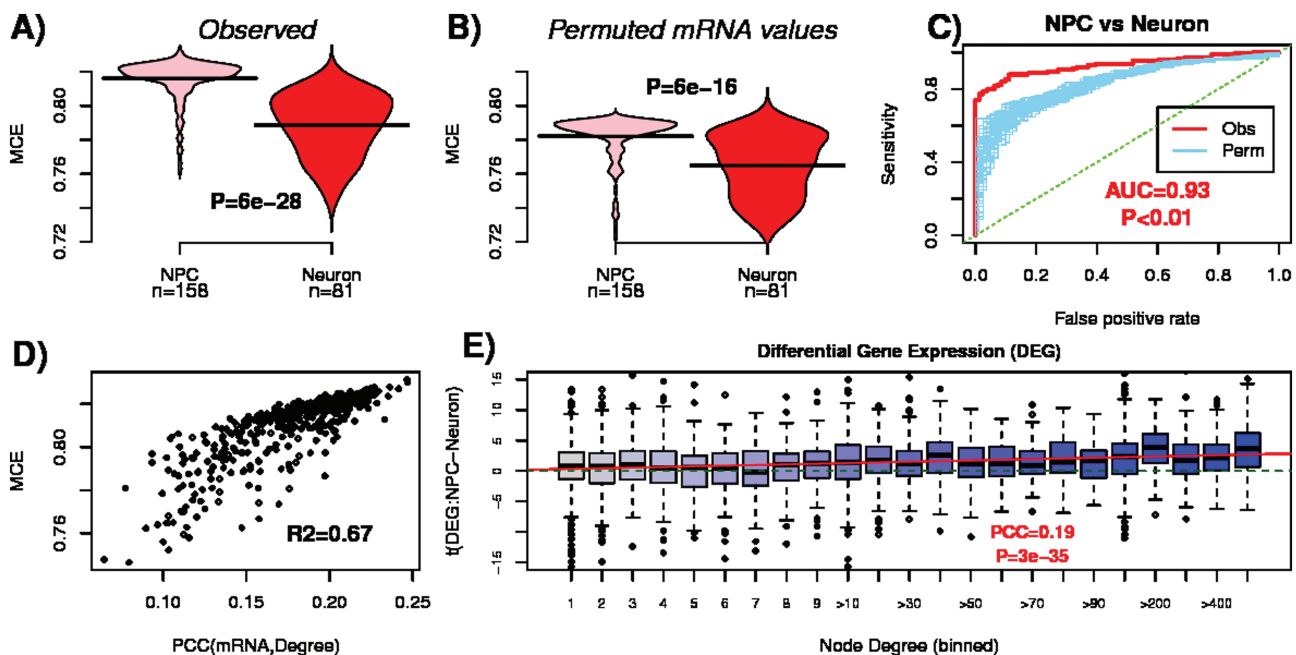
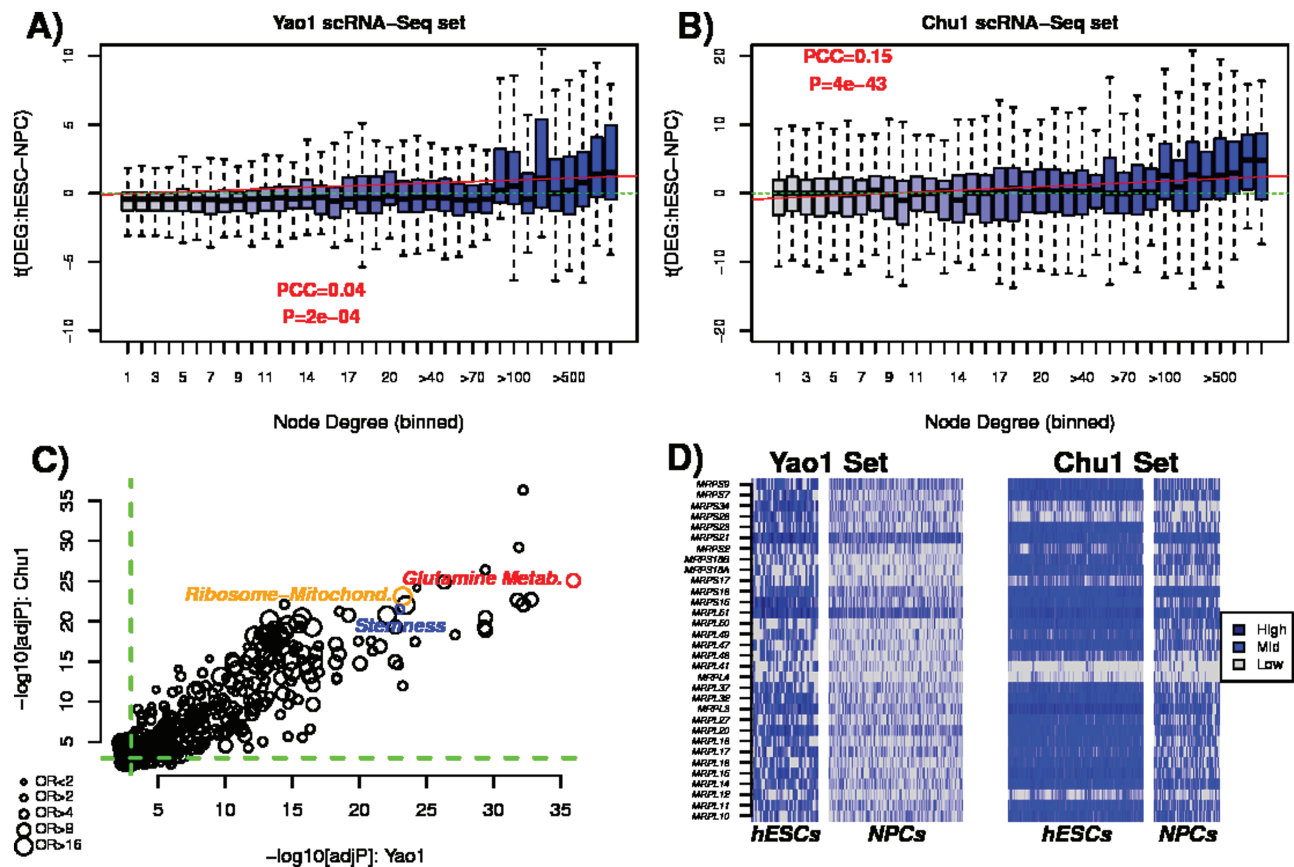


Figure 4. Integration with network information underpins the association of MCE with cell potency. (A) Violin plots of the observed MCE values for NPCs and terminally differentiated neurons (scRNA-Seq data from Wang et al.). P-value derives from a one-tailed Wilcoxon rank sum test. (B) As (A), but now for a random permutation of the expression values over the network, leading to a reduced discrimination accuracy. (C) Comparison of the AUC discrimination measure for the observed case [depicted in (A)] and for 100 distinct permutations. P-value is an empirical one comparing the observed AUC to those from the 100 permutations. (D) MCE (y-axis) is approximated reasonably well by the Pearson Correlation Coefficient (PCC) between the transcriptome and connectivity/degree profile of the network (x-axis). R^2 value is given. (E) Boxplots of t-statistics of differential expression between NPCs and neurons (y-axis) against node degree (x-axis), with larger degrees binned into equal-sized groups. Green dashed line denotes the line $t = 0$. Red line is that of a linear regression. PCC is the Pearson correlation between the t-statistics and node degree. We also give the P-value for the linear regression.

but differ significantly in terms of their construction and meaning. The transcriptomic entropy used in the StemID algorithm only uses information from the scRNA-Seq profile and approximates potency in terms of the uniformity of gene expression levels. In contrast to StemID, SLICE, SCENT and MCE use additional information to define entropy and potency. In the case of SLICE, activity estimates are obtained for clusters of genes

with similar GO annotations, by averaging expression values of genes in a given cluster, and with the entropy and potency subsequently being calculated over these cluster activity estimates. The hypothesis underlying SLICE is that a more uniform activity distribution corresponds to a more potent cell. SCENT and MCE differ from SLICE in that they don't use GO-terms but instead integrate scRNA-Seq data with a model interaction network. In



the case of SCENT, a weighted interaction network is constructed in a cell-specific manner, with the weights reflecting interaction probabilities derived from the gene expression profile of the cell. These weights define a stochastic matrix, from which a signaling entropy can then be defined, which quantifies the efficiency of diffusion of signaling over the whole network. This entropy may approximate potency if the diffusion is less efficient in more differentiated cells, which could happen as a result of activation of specific subnetworks which draws in signaling flux at the expense of diffusing to other parts of the network. While MCE also integrates the scRNA-Seq data with an interaction network, it does not specify the signaling process but infers it from the data itself using the observed gene expression profile of the sample as the invariant steady-state measure of the Markov chain. In this model, potency is the graph entropy of this inferred Markov chain process, measuring the uniformity or promiscuity of the overall process on the network. As with SCENT, the underlying hypothesis is that a more potent cell exhibits a more promiscuous signaling pattern. Although information and signaling flow is most often understood in the context of regulatory transcription factor gene target networks, which are directional in nature, mathematically, the diffusion or graph entropy measures used in SCENT/MCE are equally applicable to nondirectional networks such as the PPI networks considered here. When applying these measures to these networks

properties: (i) the relative connectivity between hubs and low-degree nodes in a network is a robust feature of such networks; (ii) big changes in gene expression, specially at hubs, will have a larger influence on the MCE/SCENT estimates, with such larger changes in mRNA expression being also more robust; and (iii) MCE/SCENT are computed genome-wide over a large number of genes, which renders them robust to the potentially large numbers of dropouts in scRNA-Seq data [10, 13].

Given that a positive correlation between transcriptome and connectome is a major driver of cell potency, it is important to discuss the nature of the factors underlying this association. One potential factor could be related to the nature or inherent biases of the PPI networks themselves [49, 50]. Although our PPI network was constructed from an integrated resource encompassing several different PPI databases (HPRD, MINT, IntAct, NCI-PID and BioGrid), a procedure known to reduce source-specific biases and the false negative rate, integration of such resources may not entirely avoid all biases. For instance, it is well known that PPI assays are more likely to identify PPIs for proteins that are highly expressed, and so if the source of PPI data is preferentially skewed toward immortalized stem-cell-like cell lines or pluripotency factors, it would follow by construction that proteins highly expressed in stem-like cells would have a higher connectivity in the resulting PPI network. Thus, stemness or potency could be hardwired in the topology of the PPI network we are considering. However, far from being a bias that needs to be adjusted for, using such networks to estimate potency can be viewed as a key advantage of our approach, as this renders the potency estimation more robust in the background of noisy scRNA-Seq data. Moreover, as we have shown here, the association with potency is driven by a number of different biological modules, some of which have been experimentally validated (in non-PPI studies) to be associated with cell potency. Indeed, our GSEA on overexpressed network hubs revealed not only biological terms associated with stemness, but also a less obvious term associated with mitochondrial ribosomal proteins. This enrichment and association were highly significant and extremely robust across independent datasets and are consistent with recent non-PPI work supporting a role for mitochondrial activity and splicing rate in influencing stemness and differentiation [20–28]. Thus, although higher expression of such proteins will favor the detection of their interactions in PPI assays performed on stem-like cells, their overexpression is also of biological significance and therefore does not necessarily constitute a factor that one would want to adjust for. Moreover, ribosomal mitochondrial proteins don't seem to have been more extensively studied at the PPI level than say pluripotency factors or cancer-related proteins [49–51], further supporting the view that the observed association of MCE/SCENT with potency is nontrivial and of novel biological significance. A more recent study has also shown that loss of potency and fate commitment in hematopoiesis is associated with a lower transcription rate of ribosomal genes [20] and that this is a conserved program across species, including higher vertebrates. Thus, while our GSEA analysis focused on the differentiation of hESCs into NPCs, it would appear that upregulation of ribosomal proteins may be a universal marker of differentiation potency, i.e. a potency marker that works for any cell type and lineage. To our knowledge, that upregulation of ribosomal proteins may be a universal marker of potency is an entirely novel biological insight gained through the use of a PPI network.

We end by noting that the observed excellent agreement between the ordering of single cells according to their differentiation potency (as predicted by MCE) and their expected potency

was achieved by estimating potency for each cell individually, starting out from a purely model-driven approach, without using any prior biological information (e.g. known marker genes, sampling time point) and without using information from other cells (i.e. no feature selection was done, nor was clustering used to group cells). This is in stark contrast to most single-cell algorithms [29–37], whose main aim is to infer cell-lineage trajectories from time-course differentiation experiments, and which all use either prior information (e.g. known marker gene expression, time point) or heuristics that lack strong biological justification, to assign cells to potency states. We propose that future single-cell analysis algorithms should focus more on developing and testing improved *in silico* models of cell potency, as this is a key step toward a more accurate and unbiased reconstruction of Waddington landscapes [5].

In summary, by performing a comprehensive analysis over 11 independent scRNA-Seq experiments, we have identified the most robust single-cell potency measures, revealing that their robustness is driven by a correlation between mRNA expression and network connectivity. This study thus provides a foundation to further improve upon the analysis of single-cell RNA-Seq data, by direct *in silico* estimation of single-cell potency.

Key Points

- Single-cell potency can be approximated by the maximum network entropy of a Markov chain signaling process on a PPI network.
- The robustness of differentiation potency measures for single cells or bulk samples improves upon integration of the scRNA-Seq profile of a cell with orthogonal network connectivity information, as provided by a PPI network.
- Differentiation potency is encoded by a positive correlation between transcriptome and connectome, with network hubs exhibiting preferential overexpression in more potent cells.
- Network hubs encoding mitochondrial ribosomal proteins exhibit graded expression levels that correlate with differentiation potency.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

National Science Foundation of China (grants 31571359, 31771464, 31401120, 91529303, 31771476, 11421101 and 91530322), Royal Society Newton Advanced Fellowship (NAF project: 522438, NAF award: 164914). The National Program on Key Basic Research Project of China (Nos. 2017YFA0505500) and the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB13040700).

References

1. Waddington CH. *Principles of Development and Differentiation*. New York: Macmillan, 1966.

2. Moris N, Pina C, Arias AM. Transition states and cell fate decisions in epigenetic landscapes. *Nat Rev Genet* 2016; **17**(11):693–703.
3. Levisky JM, Shenoy SM, Pezo RC, Singer RH. Single-cell gene expression profiling. *Science* 2002; **297**(5582):836–40.
4. Laurenti E, Gottgens B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* 2018; **553**(7689):418–26.
5. Lang AH, Li H, Collins JJ, Mehta P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. *PLoS Comput Biol* 2014; **10**(8):e1003734.
6. Grün D, Muraro MJ, Boisset JC, et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 2016; **19**(2):266–77.
7. Heinaniemi M, Nykter M, Kramer R, et al. Gene-pair expression signatures reveal lineage control. *Nat Methods* 2013; **10**(6):577–83.
8. Banerji CR, Miranda-Saavedra D, Severini S, et al. Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Sci Rep* 2013; **3**:3039.
9. Guo M, Bao EL, Wagner M, et al. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res* 2017; **45**(7):e54–e54.
10. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* 2017; **8**:15599.
11. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009; **10**(1):57–63.
12. Grün D, van Oudenaarden A. Design and analysis of single-cell sequencing experiments. *Cell* 2015; **163**(4):799–810.
13. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015; **16**(3):133.
14. Yao Z, Mich JK, Ku S, et al. A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell Stem Cell* 2017; **20**(1):120–34.
15. MacArthur BD, Lemischka IR. Statistical mechanics of pluripotency. *Cell* 2013; **154**(3):484–9.
16. Yang L, Wang WH, Qiu WL, et al. A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. *Hepatology* 2017; **66**(5):1387–1401.
17. Wang J, Jenjaroenpun P, Bhinge A, et al. Single-cell gene expression analysis reveals regulators of distinct cell subpopulations among developing human neurons. *Genome Res* 2017; **27**(11):1783–94.
18. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3**:Article3.
19. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**(43):15545–50.
20. Athanasiadis EI, Botthof JG, Andres H, et al. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat Commun* 2017; **8**(1):2045.
21. Johnston IG, Gaal B, Neves RP, et al. Mitochondrial variability as a source of extrinsic cellular noise. *PLoS Comput Biol* 2012; **8**(3):e1002416.
22. Guantes R, Rastrojo A, Neves R, et al. Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome Res* 2015; **25**(5):633–44.
23. das Neves RP, Jones NS, Andreu L, et al. Connecting variability in global transcription rate to mitochondrial variability. *PLoS Biol* 2010; **8**(12):e1000560.
24. Schieke SM, Ma M, Cao L, et al. Mitochondrial metabolism modulates differentiation and teratoma formation capacity in mouse embryonic stem cells. *J Biol Chem* 2008; **283**(42):28506–12.
25. Sukumar M, Liu J, Mehta GU, et al. Mitochondrial membrane potential identifies cells with enhanced stemness for cellular therapy. *Cell Metab* 2016; **23**(1):63–76.
26. Hu C, Fan L, Cen P, et al. Energy metabolism plays a critical role in stem cell maintenance and differentiation. *Int J Mol Sci* 2016; **17**(2):253.
27. Folmes CD, Ma H, Mitalipov S, Terzic A. Mitochondria in pluripotent stem cells: stemness regulators and disease targets. *Curr Opin Genet Dev* 2016; **38**:1–7.
28. Folmes CD, Terzic A. Energy metabolism in the acquisition and maintenance of stemness. *Semin Cell Dev Biol* 2016; **52**:68–75.
29. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014; **32**(4):381–6.
30. Marco E, Karp RL, Guo G, et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci USA* 2014; **111**(52):E5643–50.
31. Setty M, Tadmor MD, Reich-Zeliger S, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* 2016; **34**(6):637.
32. Bendall SC, Davis KL, Amir EAd, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 2014; **157**(3):714–25.
33. Chen J, Schlitzer A, Chakarov S, et al. Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat Commun* 2016; **7**.
34. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* 2017; **14**(10):979.
35. Rizvi AH, Camara PG, Kandror EK, et al. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat Biotechnol* 2017; **35**(6):551–60.
36. Haghverdi L, Büttner M, Wolf FA, et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* 2016; **13**(10):845–8.
37. Angerer P, Haghverdi L, Büttner M, et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 2015; **32**(8):1241–3.
38. Chu LF, Leng N, Zhang J, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* 2016; **17**(1):173.
39. Treutlein B, Brownfield DG, Wu AR, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single cell RNA-seq. *Nature* 2014; **509**(7500):371.
40. Semrau S, Goldmann JE, Soumilion M, et al. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat Commun* 2017; **8**(1):1096.
41. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 2016; **352**(6282):189–96.
42. Teschendorff AE, Sollich P, Kuehn R. Signalling entropy: a novel network-theoretical framework for systems analysis and interpretation of functional omic data. *Methods* 2014; **67**(3):282–93.

43. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2008;**9**(3): 432–41.
44. Mohan K, London P, Fazel M, et al. Node-based learning of multiple gaussian graphical models. *J Mach Learn Res* 2014; **15**(1):445–88.
45. Alipanahi B, Frey BJ. Network cleanup. *Nat Biotechnol* 2013; **31**(8):714.
46. Barzel B, Barabási AL. Network link prediction by global silencing of indirect correlations. *Nat Biotechnol* 2013;**31**(8): 720–5.
47. Shi J, Zhao J, Li T, Chen L. Detecting direct associations in a network by information theoretic approaches. *Sci China Math* 2017. Online <http://engine.scichina.com/publisher/scp/journal/SCM/doi/10.1007/s11425-017-9206-0?slug=abstract>.
48. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**(Nov):2579–605.
49. Gillis J, Ballouz S, Pavlidis P. Bias tradeoffs in the creation and analysis of protein-protein interaction networks. *J Proteomics* 2014;**100**:44–54.
50. Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein-protein interaction networks and biology-what's the connection? *Nat Biotech* 2008;**26**(1):69–72.
51. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 2006; **22**(18):2291–7.